**Abstract**

 Although it is not as commonly recognized as heart disease itself, chest pain can be a debilitating health problem. Additionally, the reasons for chest pain are not completely understood. This issue inspired this project, which intended to develop a model consisting of clinical measurements that can be used to predict the presence and type of chest pain in patients. This would also shine light onto the causal factors of chest pain. The nominal types of chest pain investigated include typical angina, atypical angina, and nonanginal pain. Based on our analysis, the presence of exercise-induced angina, ST depression induced by exercise relative to rest, and the presence of heart disease were determined as significant predictors. Odds ratios for each chest pain type were also determined. In regard to the objective of creating a predictive model, we were only moderately successful as the model had an error rate of 41.6%.

**Introduction**

Heart disease is one of the leading causes of death in the world. According to the CDC, it is projected that by 2030 the United States alone will spend more than $818 billion dollars on medical costs associated with cardiovascular disease. However, as is the case with many diseases, many of the costs of heart disease cannot be measured monetarily but are rather present in the pain and inconvenience felt by patients. For this reason, it is important that the scientific community learn more about the reasons for chest pain so that causal factors and mechanisms of chest pain can be determined. This project attempts to investigate this issue and determine predictors for different types of chest pain. Potential benefits of this investigation include learning reasons for chest pain while also providing a predictive model for what kind of chest pain a person is likely to feel based off of their characteristics.

**Problem Statement**

As previously described, heart disease is associated with huge costs in our society. Beyond the financial burden, patients must live with pain that can affect their life immensely. Researchers have attempted to collect data on patients surrounding this subject. The data used in this investigation was collected from patients who were treated at Cleveland Hospital, and is now kept in the University of California, Irvine Machine Learning Repository. One large obstacle in this area is that there are different types of chest pain that are not necessarily ordinal. Instead, they are more nominal in nature. Unlike a rating involving levels of pain, chest pain is commonly split up into three categories: typical angina, atypical angina, and nonanginal pain. Table 1 in the appendix shows the clinical differences between chest pain types. This investigation seeks to answer how the type of chest pain can be predicted from patient characteristics and values of commonly performed medical tests.

**Goals and Objectives**

The purpose of our project is to find which factors have a significant effect on type of chest pain in patients, and what that effect is. With this knowledge, we hope to create a model that can be used to accurately predict the type of chest pain in patients. The objectives can be summed up as answering the questions of what factors are significant in predicting presence and type of chest pain, and can a model be created to accurately predict type of chest pain in patients given their medical characteristics?

 There are many potential benefits of this study. First, we hope to learn about the most significant factors that play a role in chest pain. With this knowledge, physicians can first focus on these factors in their patients before moving on to other tests. Also, a model could be used to give a likelihood and determine if it is “normal” for a person to be feeling this type of chest pain. By determining a probability of a person feeling a certain type of chest pain based on certain factors, it can give a physician an idea if it is likely that there are other factors at play in the patient’s chest pain. Finally, a model could possibly be used from a public health perspective to predict the likelihood and prevalence of people experiencing certain types of chest pain.

**Background**

Our data was collected from the Cleveland Health database and has been used by machine learning researchers. The dataset includes 13 predictor variables that are either binary, ordinal, or continuous and our response variable (chest pain). Table 2 in the appendix describes the variables. As previously stated, chest pain is not rating of pain but is split up into 4 categories: no pain, typical angina, atypical angina, and nonanginal pain. Figure 1 in the appendix displays the distribution of chest pain and shows that most patients in this dataset had no chest pain.

Within our predictors we suspect that there may be multicollinearity between factors. We discovered that presence of heart disease correlates highly with many variables such as chest pain, presence of exercise induced angina, and ST depression induced by exercise relative to rest. Since the data are categorical, the first option that one would think of to analyze the data is ordinary logistic regression. However, ordinary logistic regression would not be the best option in the case because the outcome isn't binary and there are instead five different possible outcomes. Another option would be other models such as the cumulative logit model that can be used for ordinal categories. However, this is also not ideal since the data are nominal rather than ordinal. Since the data are nominal, a generalized logit model would be a more appropriate choice for this investigation. This model works by modeling response functions for each response that contrast with each other.

**Methodology**

As stated, we will use model type of chest pain with a generalized logit model within a proc logistic step in SAS. To choose significant variables to be included in the model, we will use forward selection methods to create a model. In order to identify potential interactions between variables, two-factor interactions will be eligible to be included in the model. Once the model has been created, we will be able to calculate parameters that will allow us to predict the likelihood of each type of chest pain. Then, we will use a hypothetical patient to demonstrate how the model can be used to calculate the probability of a patient experiencing a certain type of chest pain. We will also obtain odds ratios between levels of predictor variables to show the effect of each variable.

 Limitations of this methodology include the fact that the model is not guaranteed to be made up of causal predictors of chest pain and instead may reflect a correlation. However, hopefully the possibility of including interaction terms will help tease apart this issue. Also, the investigation into collinearity will help us detect underlying trends and confounders that may play a part in chest pain.

**Analysis and Results**

After using forward selection methods to make the model, the presence of exercise-induced angina, the presence of heart disease, and the ST depression induced by exercise relative to rest were all included as significant variables in each of the models. Parameters are shown in Table 3 and the odds ratios between levels of predictors are shown in Table 4. The odds ratios are useful in comprehending the effects of each predictor. For example, those with heart disease are 6.3 times as likely to experience non-anginal pain than those without heart disease. The parameters for each pain type can be plugged into the model to determine the probability of a patient experiencing a certain type of chest pain as shown by Equation 1. To use the equation, one must simply use the $β$-coefficients for the chest pain type to be calculated, as given by Table 3.

$$lnPr⁡(presence)/Pr⁡(no presence) =β\\_o+β\\_1 (exercise-induced angina)+β\\_2 (ST depression)+β\\_3 (heart disease)$$

*Equation 1. Chest pain type probabilities*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CP | Intercept ($β$0)  | Presence of exercise induced angina ($β$1)  | ST depression induced by exercise ($β$2)  | Presence of heart disease ($β$3)  |
| 1 | -0.7485 | -1.9476 | -0.9762 | 1.5445 |
| 2 | -0.9643 | -1.5876 | -0.0442 | 1.8610 |
| 3 | -2.7688 | -1.4177 | 0.3923 | 1.8424 |

*Table 3. Parameters relating to each chest pain type equation*

|  |  |  |  |
| --- | --- | --- | --- |
| CP | Presence of exercise-induced angina 1 vs 0 odds ratio | ST depression induced by exercise odds ratio  | Presence of heart disease 1 vs 0 odds ratio |
| 1 | 0.143 | 0.377 | 4.686 |
| 2 | 0.204 | 0.957 | 6.430 |
| 3 | 0.242 | 1.480 | 6.312 |

*Table 4. Odds ratios of each parameter relating to each chest pain type equation*

After creating the model, the model was used to predict each patient's likelihood of experiencing a certain type of chest pain. A hypothetical patient with the presence of exercise-induced angina, an ST depression induced by exercise value of 1.0, and the presence of heart disease was also created and the probabilities of the patient experiencing a certain type of chest pain were calculated. The patient was predicted to have a 57.5% probability of experiencing no chest pain, a 6.8% probability of experiencing typical angina, a 27.5% probability of experiencing atypical angina, and an 8.2% probability of experiencing nonanginal pain.

Finally, significance of variables across chest pain types and fit statistics were assessed. Some of the terms, such as the ST depression induced by exercise term in the atypical angina model, were not significant in the models for every chest pain type. This means that perhaps some types of chest pain are more strongly predicted by certain variables. Additionally, the model had an error rate of 41.6%. This means that 41.6% of the patients were misclassified by the model. Although this is not a very good rate of successful classification, it is still significantly better than choosing at random, which shows the predictive power of the model.

**Findings and Conclusion**

 This study sought to determine the variables relevant to predicting the presence and type of chest pain. The presence of exercise-induced angina, the presence of heart disease, and the ST depression induced by exercise relative to rest were all determined to have significant effects on the presence of certain types of chest pain, and their parameters show their relative influence. Especially with further refinements, the model created could be used in many applications. For example, it could be used to predict the likelihood of patients experiencing certain types of chest pain. This could help physicians reason if it is expected for a patient to be feeling a certain type of chest pain and determine if there are underlying factors at play. Additionally, the inclusion of certain variables as relevant to chest pain type could be used to help find causes of chest pain and thus potential treatments.

Regarding the usefulness of the model, his study was somewhat successful. Although a useful model was developed, the model still had an error rate of 41.6%. As is that case with most medical tests, it is essential to be accurate and precise in modeling to avoid the emotional pain and monetary costs that come with false diagnoses and additional testing. As such, the model may not be suitable for its intended purpose. However, it does have some value and could be used as a basis for building a more useful model.

 Future work could also build upon the objectives of this study. For example, the association of types of chest pain with traditional outcomes such as mortality could be determined so that different types of negative outcomes can be compared. Also, it is possible that there are common causes of both chest pain and our variables, such as potentially genetics, that were not measured in our investigation. Further investigation into causes of heart disease and pain may turn out to have more predictive power, but data concerning these relationships may be more difficult to obtain.

**Appendix**

|  |  |
| --- | --- |
| Typical Angina | Meets three of the following characteristics:* Substernal chest discomfort of characteristic quality and duration
* Provoked by exertion or emotional stress
* Relieved by rest and/or nitroglycerine
 |
| Atypical Angina | Meets two of following characteristics.* Substernal chest discomfort of characteristic quality and duration
* Provoked by exertion or emotional stress
* Relieved by rest and/or nitroglycerine
 |
| Nonanginal Pain | * Duration is over 30 minutes or less than 5 seconds
* It increases with inspiration, can be brought on with one movement of the trunk or arm, can be brought on by local fingers pressure, or bending forward,
* It can be relieved immediately on lying down.
 |

*Table 1. Clinical Classification of Chest Pain*



 *Figure 1. Distribution of Chest Pain*



*Table 2. Variable descriptions and levels*

**References**

<https://www.cdcfoundation.org/pr/2015/heart-disease-and-stroke-cost-america-nearly-1-billion-day-medical-costs-lost-productivity>

<https://www.kaggle.com/ronitf/heart-disease-uci>